PATENT & TRADEMARK OFFICE
FEB 07 2002
JC95 3E

Docket No. 10806-152 **PATENT**

CERTIFICATE OF MAILING

I hereby certify that this paper is being deposited with the United States Postal Service with sufficient postage as first class mail in an envelope addressed to: Assistant Commissioner of Patents, Washington, DC 20231 on Jan. 24 2002.

Laureen E Mans

IN THE UNITED STATES PATENT & TRADEMARK OFFICE

Applicant:	Anneli Attersand	:	Paper No.:	
Serial No.:	09/990,415	:	Group Art Unit:	1614
Filing Date:	November 21, 2001	:	Examiner:	

For: **Gene Encoding Protein Cluster I and the Encoded Protein**

TRANSMITTAL OF CERTIFIED PRIORITY DOCUMENT

Assistant Commissioner for Patents
Washington, D.C. 20231

Dear Sir:

Submitted herewith is a certified copy of Swedish priority application No. 0004325-7 filed November 24, 2000. The priority document is in the English language. It is believed that this satisfies the requirements of 35 U.S.C. §119.

Respectfully submitted,

By _____
Holly D. Kozlowski, Reg. No. 30,468
Attorney for Applicant
DINSMORE & SHOHL LLP
1900 Chemed Center
255 East Fifth Street
Cincinnati, Ohio 45202
(513) 977-8568

750749v1

PRV

PATENT- OCH REGISTRERINGSVERKET
Patentavdelningen

Intyg
Certificate

Härmed intygas att bifogade kopior överensstämmer med de handlingar som ursprungligen ingivits till Patent- och registreringsverket i nedannämnda ansökan.

This is to certify that the annexed is a true copy of the documents as originally filed with the Patent- and Registration Office in connection with the following patent application.

(71) *Sökande* / *Applicant (s)* — *Pharmacia AB, Stockholm SE*

(21) *Patentansökningsnummer* / *Patent application number* — *0004325-7*

(86) *Ingivningsdatum* / *Date of filing* — *2000-11-24*

Stockholm, 2001-11-21

För Patent- och registreringsverket
For the Patent- and Registration Office

Hjördis Segerlund

Hjördis Segerlund

Avgift
Fee 170:-

PATENT- OCH REGISTRERINGSVERKET
SWEDEN

Postadress/Adress
Box 5055
S-102 42 STOCKHOLM

Telefon/Phone
+46 8 782 25 00
Vx 08-782 25 00

Telex
17978
PATOREG S

Telefax
+46 8 666 02 86
08-666 02 86

PROTEIN CLUSTER I

TECHNICAL FIELD

The present invention relates to the identification of a human gene family expressed in metabolically relevant tissues. The genes encode a group polypeptides referred to as "Protein Cluster I" which are predicted to be useful in the diagnosis of metabolic diseases, such as obesity and diabetes, as well as in the identification of agents useful in the treatment of the said diseases.

BACKGROUND ART

Metabolic diseases are defined as any of the diseases or disorders that disrupt normal metabolism. They may arise from nutritional deficiencies; in connection with diseases of the endocrine system, the liver, or the kidneys; or as a result of genetic defects. Metabolic diseases are conditions caused by an abnormality in one or more of the chemical reactions essential to producing energy, to regenerating cellular constituents, or to eliminating unneeded products arising from these processes. Depending on which metabolic pathway is involved, a single defective chemical reaction may produce consequences that are narrow, involving a single body function, or broad, affecting many organs and systems.

One of the major hormones that influence metabolism is insulin, which is synthesized in the beta cells of the islets of Langerhans of the pancreas. Insulin primarily regulates the direction of metabolism, shifting many processes toward the storage of substrates and away from their degradation. Insulin acts to increase the transport of glucose and amino acids as well as key minerals such as potassium, magnesium, and phosphate from the blood into cells. It also regulates a variety of enzymatic reactions within the cells, all of which have a common overall direction, namely the synthesis of large molecules from small units. A deficiency in the action of insulin (diabetes mellitus) causes severe impairment in (i) the storage of glucose in the form of glycogen and the oxidation of

glucose for energy; (ii) the synthesis and storage of fat from fatty acids and their precursors and the completion of fatty-acid oxidation; and (iii) the synthesis of proteins from amino acids.

There are two varieties of diabetes. Type I is insulin-dependent diabetes mellitus (IDDM), for which insulin injection is required; it was formerly referred to as juvenile onset diabetes. In this type, insulin is not secreted by the pancreas and hence must be taken by injection. Type II, non-insulin-dependent diabetes mellitus (NIDDM) may be controlled by dietary restriction. It derives from insufficient pancreatic insulin secretion and tissue resistance to secreted insulin, which is complicated by subtle changes in the secretion of insulin by the beta cells. Despite their former classifications as juvenile or adult, either type can occur at any age; NIDDM, however, is the most common type, accounting for 90 percent of all diabetes. While the exact causes of diabetes remain obscure, it is evident that NIDDM is linked to heredity and obesity. There is clearly a genetic predisposition to NIDDM diabetes in those who become overweight or obese.

Obesity is usually defined in terms of the body mass index (BMI), i.e. weight (in kilograms) divided by the square of the height (in meters). Weight is regulated with great precision. Regulation of body weight is believed to occur not only in persons of normal weight but also among many obese persons, in whom obesity is attributed to an elevation in the set point around which weight is regulated. The determinants of obesity can be divided into genetic, environmental, and regulatory.

Recent discoveries have helped explain how genes may determine obesity and how they may influence the regulation of body weight. For example, mutations in the *ob* gene have led to massive obesity in mice. Cloning the *ob* gene led to the identification of leptin, a protein coded by this gene; leptin is produced in adipose tissue cells and acts to control body fat. The existence of leptin supports the idea that body weight is regulated, because leptin serves as a signal between adipose tissue and the areas of the brain that control energy metabolism, which influences body weight.

Metabolic diseases like diabetes and obesity are clinically and genetically heterogeneous disorders. Recent advances in molecular genetics have led to the recognition of genes involved in IDDM and in some subtypes of NIDDM, including maturity-onset diabetes of the young (MODY) (Velho & Froguel (1997) Diabetes Metab. 23 Suppl 2:34-37). However, several IDDM susceptibility genes have not yet been identified, and very little is known about genes contributing to common forms of NIDDM. Studies of candidate genes and of genes mapped in animal models of IDDM or NIDDM, as well as whole genome scanning of diabetic families from different populations, should allow the identification of most diabetes susceptibility genes and of the molecular targets for new potential drugs. The identification of genes involved in metabolic disorders will thus contribute to the development of novel predictive and therapeutic approaches.

BRIEF DESCRIPTION OF THE DRAWING

Fig. 1

Transmembrane regions identified in the proteins shown as (a) SEQ ID NO: 2; (b) SEQ ID NO: 8; and (c) SEQ ID NO: 6.

DESCRIPTION OF THE INVENTION

According to the present invention, a family of genes and encoded homologous proteins (hereinafter referred to as "Protein Cluster I") has been identified. Consequently, the present invention provides an isolated nucleic acid molecule selected from:

(a) nucleic acid molecules comprising a nucleotide sequence as shown in SEQ ID NO: 1, 3, 5 or 7;

(b) nucleic acid molecules comprising a nucleotide sequence capable of hybridizing, under stringent hybridization conditions, to a nucleotide sequence complementary to the polypeptide coding region of a nucleic acid molecule as defined in (a); and

(c) nucleic acid molecules comprising a nucleic acid sequence which is degenerate as a result of the genetic code to a nucleotide sequence as defined in (a) or (b).

The nucleic acid molecules according to the present invention includes cDNA, chemically synthesized DNA, DNA isolated by PCR, genomic DNA, and combinations thereof. RNA transcribed from DNA is also encompassed by the present invention.

The term "stringent hybridization conditions" is known in the art from standard protocols (e.g. Ausubel et al., *supra*) and could be understood as e.g. hybridization to filter-bound DNA in 0.5 M $NaHPO_4$, 7% sodium dodecyl sulfate (SDS), 1 mM EDTA at +65°C, and washing in 0.1xSSC / 0.1% SDS at +68°C.

In a preferred form of the invention, the said nucleic acid molecule has a nucleotide sequence identical with SEQ ID NO: 1 of the Sequence Listing. However, the nucleic acid molecule according to the invention is not to be limited strictly to the sequence shown as SEQ ID NO: 1. Rather the invention encompasses nucleic acid molecules carrying modifications like substitutions, small deletions, insertions or inversions, which nevertheless encode proteins having substantially the features of the Protein Cluster I polypeptide according to the invention. Included in the invention are consequently nucleic acid molecules, the nucleotide sequence of which is at least 90% homologous, preferably at least 95% homologous, with the nucleotide sequence shown as SEQ ID NO: 1 in the Sequence Listing.

Included in the invention is also a nucleic acid molecule which nucleotide sequence is degenerate, because of the genetic code, to the nucleotide sequence shown as SEQ ID NO: 1. A sequential grouping of three nucleotides, a "codon", codes for one amino acid. Since there are 64 possible codons, but only 20 natural amino acids, most amino acids are coded for by more than one codon. This natural "degeneracy", or "redundancy", of the genetic code is well known in the art. It will thus be appreciated that the nucleotide sequence shown in the Sequence Listing is only an example within a large but definite group of sequences which will encode the Protein Cluster I polypeptide.

The nucleic acid molecules according to the invention have numerous applications in techniques known to those skilled in the art of molecular biology. These techniques include their use as hybridization probes, for chromosome and gene mapping, in PCR technologies, in the production of sense or antisense nucleic acids, in screening for new therapeutic molecules, etc.

More specifically, the sequence information provided by the invention makes possible large-scale expression of the encoded polypeptides by techniques well known in the art. Nucleic acid molecules of the invention also permit identification and isolation of nucleic acid molecules encoding related polypeptides, such as human allelic variants and species homologues, by well-known techniques including Southern and/or Northern hybridization, and PCR. Knowledge of the sequence of a human DNA also makes possible, through use of Southern hybridization or PCR, the identification of genomic DNA sequences encoding the proteins in Cluster I, expression control regulatory sequences such as promoters, operators, enhancers, repressors, and the like. Nucleic acid molecules of the invention are also useful in hybridization assays to detect the capacity of cells to express the proteins in Cluster I. Nucleic acid molecules of the invention may also provide a basis for diagnostic methods useful for identifying a genetic alteration(s) in a locus that underlies a disease state or states, which information is useful both for diagnosis and for selection of therapeutic strategies.

In a further aspect, the invention provides an isolated polypeptide encoded by the nucleic acid molecule as defined above. In a preferred form, the said polypeptide has an amino acid sequence according to SEQ ID NO: 2, 4, 6 or 8 of the Sequence Listing. However, the polypeptide according to the invention is not to be limited strictly to a polypeptide with an amino acid sequence identical with SEQ ID NO: 2, 4, 6 or 8 in the Sequence Listing. Rather the invention encompasses polypeptides carrying modifications like substitutions, small deletions, insertions or inversions, which polypeptides nevertheless have substantially the features of the Protein Cluster I polypeptide. Included in the invention are consequently polypeptides, the amino acid sequence of which is at least 90% homologous, preferably at least 95% homologous,

with the amino acid sequence shown as SEQ ID NO: 2, 4, 6 or 8 in the Sequence Listing.

In a further aspect, the invention provides a vector harboring the nucleic acid molecule as defined above. The said vector can e.g. be a replicable expression vector, which carries and is capable of mediating the expression of a DNA molecule according to the invention. In the present context the term "replicable" means that the vector is able to replicate in a given type of host cell into which is has been introduced. Examples of vectors are viruses such as bacteriophages, cosmids, plasmids and other recombination vectors. Nucleic acid molecules are inserted into vector genomes by methods well known in the art.

Included in the invention is also a cultured host cell harboring a vector according to the invention. Such a host cell can be a prokaryotic cell, a unicellular eukaryotic cell or a cell derived from a multicellular organism. The host cell can thus e.g. be a bacterial cell such as an *E. coli* cell; a cell from a yeast such as *Saccharomyces cervisiae* or *Pichia pastoris*, or a mammalian cell. The methods employed to effect introduction of the vector into the host cell are standard methods well known to a person familiar with recombinant DNA methods.

In yet another aspect, the invention provides a process for production of a polypeptide, comprising culturing a host cell, according to the invention, under conditions whereby said polypeptide is produced, and recovering said polypeptide. The medium used to grow the cells may be any conventional medium suitable for the purpose. A suitable vector may be any of the vectors described above, and an appropriate host cell may be any of the cell types listed above. The methods employed to construct the vector and effect introduction thereof into the host cell may be any methods known for such purposes within the field of recombinant DNA. The recombinant polypeptide expressed by the cells may be secreted, i.e. exported through the cell membrane, dependent on the type of cell and the composition of the vector.

In a further aspect, the invention provides a method for identifying an agent capable of modulating a nucleic acid molecule according to the invention, comprising
(i) providing a cell comprising the said nucleic acid molecule;
(ii) contacting said cell with a candidate agent; and
(iii) monitoring said cell for an effect that is not present in the absence of said candidate agent.

For screening purposes, appropriate host cells can be transformed with a vector having a reporter gene under the control of the nucleic acid molecule according to this invention. The expression of the reporter gene can be measured in the presence or absence of an agent with known activity (i.e. a standard agent) or putative activity (i.e. a "test agent" or "candidate agent"). A change in the level of expression of the reporter gene in the presence of the test agent is compared with that effected by the standard agent. In this way, active agents are identified and their relative potency in this assay determined.

A transfection assay can be a particularly useful screening assay for identifying an effective agent. In a transfection assay, a nucleic acid containing a gene such as a reporter gene that is operably linked to a nucleic acid molecule according to the invention, is transfected into the desired cell type. A test level of reporter gene expression is assayed in the presence of a candidate agent and compared to a control level of expression. An effective agent is identified as an agent that results in a test level of expression that is different than a control level of reporter gene expression, which is the level of expression determined in the absence of the agent. Methods for transfecting cells and a variety of convenient reporter genes are well known in the art (see, for example, Goeddel (ed.), Methods Enzymol., Vol. 185, San Diego: Academic Press, Inc. (1990); see also Sambrook, *supra*).

Throughout this description the terms "standard protocols" and "standard procedures", when used in the context of molecular biology techniques, are to be understood as protocols and procedures found in an ordinary laboratory manual such as: Current Protocols in Molecular Biology, editors F. Ausubel et al., John Wiley and Sons, Inc.

1994, or Sambrook, J., Fritsch, E.F. and Maniatis, T., Molecular Cloning: A laboratory manual, 2nd Ed., Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY 1989.

Additional features of the invention will be apparent from the following Examples. Examples 1 to 3 are actual, while Examples 4 to 9 are prophetic.

EXAMPLES

EXAMPLE 1: Identification of protein clusters

A family of homologous proteins (hereinafter referred to as "Protein Cluster I") was identified by an "all-versus-all" BLAST procedure using all *Caenorhabditis elegans* proteins in the Wormpep20 database release (*http://www.sanger.ac.uk/Projects/C_elegans/wormpep/index.shtml*). The Wormpep database contains the predicted proteins from the *C. elegans* genome sequencing project, carried out jointly by the Sanger Centre in Cambridge, UK and the Genome Sequencing Center in St. Louis, USA. A number of 18,940 proteins were retrieved from Wormpep20. The proteins were used in a Smith-Waterman clustering procedure to group together proteins of similarity (Smith T.F. & Waterman M.S. (1981) *Identification of common molecular subsequences.* J. Mol. Biol. 147(1): 195-197; Pearson WR. (1991) *Searching protein sequence libraries: comparison of the sensitivity and selectivity of the Smith-Waterman and FASTA algorithms.* Genomics 11: 635-650; Olsen et al. (1999) *Optimizing Smith-Waterman alignments.* Pac Symp Biocomput.302-313). Completely annotated proteins were filtered out, whereby 10,130 proteins of unknown function could be grouped into 1,800 clusters.

The obtained sequence clusters were compared to the *Drosophila melanogaster* proteins contained in the database Flybase (Berkeley Drosophila Genome Project; *http://www.fruitfly.org*), and annotated clusters were removed. Non-annotated protein clusters, conserved in both *C. elegans* and *D. melanogaster*, were saved to a worm/fly data set, which was used in a BLAST procedure (*http://www.ncbi.nlm.nih.gov/*

Education/BLASTinfo/information3.html) against the Celera Human Genome Database (*http://www.celera.com*). Overlapping fragments were assembled to, as close as possible, full-length proteins using the PHRAP software, developed at the University of Washington *(http://www.genome.washington.edu/ UWGC/analysistools/phrap.htm)*. A group of homologous proteins ("Protein Cluster I") with unknown function was chosen for further studies.

EXAMPLE 2: Analyses of Protein Cluster I

(a) Alignment

The human part of Protein Cluster I comprises polypeptides encoded by three genes (SEQ ID NOS: 1, 5 and 7). In addition, an alternative splicing (corresponding to a deletion of positions 624 to 794 of the gene shown as SEQ ID NO: 1 results in SEQ ID NO: 3. The gene shown as SEQ ID NO: 1 was found to be comprised in a human DNA sequence from clone RP11-108L7 on chromosome 10 (GenBank Accession No. AL133215).

An alignment of the human polypeptides included in Protein Cluster I (SEQ ID NOS: 2, 4, 6 and 8), using the ClustalX multiple alignment software (downloadable from e.g. *ftp://ftp.ebi.ac.uk*) is shown in Table I. For references to the ClustalX software, see Thompson et al. (1997) *The ClustalX windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools*; Nucleic Acids Research, 24:4876-4882. See also Jeanmougin et al. (1998) *Multiple sequence alignment with ClustalX*; Trends Biochem. Sci. 23:403-405. The alignment showed a high degree of conservation in two separate regions, indicating the presence of two novel domains (see positions marked with stars in Table I).

(b) HMM-Pfam

A HMM-Pfam search was performed on the three human family members. Pfam (*http://pfam.wustl.edu*) is a large collection of protein families and domains. Pfam contains multiple protein alignments and profile-HMMs (Profile Hidden Markov Models) of these families. Profile-HMMs can be used to do sensitive database searching using statistical descriptions of a sequence family's consensus. Pfam is available on the WWW at *http://pfam.wustl.edu; http://www.sanger.ac.uk/Software/Pfam*; and *http://www.cgr.ki.se/Pfam*. The latest version (4.3) of Pfam contains 1815 families. These Pfam families match 63% of proteins in SWISS-PROT 37 and TrEMBL 9. For references to Pfam, see Bateman et al. (2000) *The Pfam protein families database*. Nucleic Acids Res. 28:263-266; Sonnhammer et al. (1998) *Pfam: Multiple Sequence Alignments and HMM-Profiles of Protein Domains*. Nucleic Acids Research, 26:322-325; Sonnhammer et al. (1997) *Pfam: a Comprehensive Database of Protein Domain Families Based on Seed Alignments*. Proteins 28:405-420.

The HMM-Pfam search indicated that no previously known domains could be identified in Protein Cluster I.

(c) TM-HMM

The human proteins in Cluster I were analyzed using the TM-HMM tool available at *http://www.cbs.dtu.dk/services/TMHMM-1.0*. TM-HMM is a method to model and predict the location and orientation of alpha helices in membrane-spanning proteins (Sonnhammer et al. (1998) *A hidden Markov model for predicting transmembrane helices in protein sequences*. ISMB 6:175-182). Transmembrane segments were identified in the proteins shown as SEQ ID NOS: 2, 6 and 8 (Fig. 1)

(d) Analysis of non-human orthologs

The *C. elegans* genome includes six genes encoding proteins within Protein Cluster I, of which the closest ancestor in evolution, a sequence included the *C. elegans* cosmid

T04F8.1 (GenBank Accession No. Z66565; see also: *Genome sequence of the nematode C. elegans: a platform for investigating biology;* The *C. elegans* Sequencing Consortium. Science (1998) 282:2012-2018. Published errata appear in Science (1999) 283:35; 283:2103; and 285:1493.) is 53% identical to the three identified human proteins (SEQ ID NOS: 2, 6 and 8).

The *Drosophila melanogaster* genome comprises two genes belonging to Protein Cluster I, of which the closest relative (GenBank Accession No. AE003606_24; see also Adams et al. (2000) *The genome sequence of Drosophila melanogaster*; Science 287:2185-2195) is 53% identical to the human protein set.

The human proteins also show 38% identity to a *Saccharomyces cerevisiae* protein (GenPept Accession No. CAA99495.1). The yeast protein has been annotated in the *Saccharomyces* Genome Database as a putative transporter (*http://genome-www4.stanford.edu/cgi-bin/SGD/locus.pl?locus=YOR270c*).

Two public *Rattus norvegicus* database entries (GENBANK entries AF276997 and S70011) have been annotated as putative tricarboxylate carrier proteins. The genes have 88% and 79% identity, respectively, with SEQ ID NO: 1. The tricarboxylate carrier transports citrate or other tricarboxylates across the inner membranes of mitochondria in an electroneutral exchange for malate or other dicarboxylic acids. (Azzi et al. (1993) J. Bioenerg. Biomembr. 25: 515-524).

EXAMPLE 3: Expression analysis

EST databases provided by the EMBL (*http://www.embl.org/Services/index.html*) were used to check whether the human proteins in Cluster I were expressed, in order to identify putative pseudogenes. No putative pseudogenes were identified in Protein Cluster I.

The tissue distribution of the human genes was studied using the Incyte LifeSeq® database (*http://www.incyte.com*). The nucleic acid molecule shown as SEQ ID NO: 1 was found to be expressed primarily in the nervous system and the digestive system. The nucleic acid molecule shown as SEQ ID NO: 3 was expressed primarily in male genitalia. The nucleic acid molecule shown as SEQ ID NO: 5 was expressed primarily in the liver and in embryonic structures. The nucleic acid molecule shown as SEQ ID NO: 7 was expressed primarily in the immune system. Therefore, the said nucleic acid molecules shown as SEQ ID NO: 1, 3, 5 and 7 and the polypeptides shown as SEQ ID NO: 2, 4, 6 and 8 are proposed to be useful for differential identification of the tissue(s) or cell types(s) present in a biological sample and for diagnosis of diseases and disorders, including metabolic disorders and immune disorders.

EXAMPLE 4: Multiple Tissue Northern blotting

Multiple Tissue Northern blotting (MTN) is performed to make a more thorough analysis of the expression profiles of the proteins in Cluster I. Multiple Tissue Northern (MTN™) Blots *(http://www.clontech.com/mtn)* are pre-made Northern blots featuring Premium Poly A+ RNA from a variety of different human, mouse, or rat tissues. MTN Blots can be used to analyze size and relative abundance of transcripts in different tissues. MTN Blots can also be used to investigate gene families and alternate splice forms and to assess cross species homology.

EXAMPLE 5: Expressing profiling using microarrays

Microarrays consist of a highly ordered matrix of thousands of different DNA sequences that can be used to measure DNA and RNA variation in applications that include gene expression profiling, comparative genomics and genotyping (For recent reviews, see e.g.: Harrington et al. (2000) *Monitoring gene expression using DNA microarrays*. Curr. Opin. Microbiol. 3(3): 285-291; or Duggan et al. (1999) *Expression profiling using cDNA Microarrays*. Nature Genetics Supplement 21:10-14).

The expression pattern of the proteins in Cluster I can be analyzed using GeneChip® expression arrays (*http://www.affymetrix.com/products/app_exp.html*). Briefly, mRNAs are extracted from various tissues. They are reverse transcribed using a T7-tagged oligo-dT primer and double-stranded cDNAs are generated. These cDNAs are then amplified and labeled using In Vitro Transcription (IVT) with T7 RNA polymerase and biotinylated nucleotides. The populations of cRNAs obtained are purified and fragmented by heat to produce a distribution of RNA fragment sizes from approximately 35 to 200 bases. GeneChip® expression arrays are hybridized with the samples. The arrays are washed and stained. The cartridges are scanned using a confocal scanner and the images are analyzed with the GeneChip 3.1 software (Affymetrix).

EXAMPLE 6: Identification of polypeptides binding to Protein Cluster I

In order to assay for proteins interacting with Protein Cluster I, the two-hybrid screening method can be used. The two-hybrid method, first described by Fields & Song (1989) Nature 340:245-247, is a yeast-based genetic assay to detect protein-protein interactions *in vivo*. The method enables not only identification of interacting proteins, but also results in the immediate availability of the cloned genes for these proteins.

The two-hybrid method can be used to determine if two known proteins (i.e. proteins for which the corresponding genes have been previously cloned) interact. Another important application of the two-hybrid method is to identify previously unknown proteins that interact with a target protein by screening a two-hybrid library. For reviews, see e.g.: Chien et al. (1991) *The two-hybrid system: a method to identify and clone genes for proteins that interact with a protein of interest*. Proc. Natl. Acad. Sci. U.S.A. 88:9578-9582; Bartel PL, Fields (1995) *Analyzing protein-protein interactions using two-hybrid system*. Methods Enzymol. 254:241-263; or Wallach et al. (1998) *The yeast two-hybrid screening technique and its use in the study of protein-protein*

interactions in apoptosis. Curr. Opin. Immunol. 10(2): 131-136. See also *http://www.clontech.com/matchmaker*.

The two-hybrid method uses the restoration of transcriptional activation to indicate the interaction between two proteins. Central to this technique is the fact that many eukaryotic transcriptional activators consist of two physically discrete modular domains: the DNA-binding domain (DNA-BD) that binds to a specific promoter sequence and the activation domain (AD) that directs the RNA polymerase II complex to transcribe the gene downstream of the DNA binding site. The DNA-BD vector is used to generate a fusion of the DNA-BD and a bait protein X, and the AD vector is used to generate a fusion of the AD and another protein Y. An entire library of hybrids with the AD can also be constructed to search for new or unknown proteins that interact with the bait protein. When interaction occurs between the bait protein X and a candidate protein Y, the two functional domains, responsible for DNA binding and activation, are tethered, resulting in functional restoration of transcriptional activation. The two hybrids are cotransformed into a yeast host strain harboring reporter genes containing appropriate upstream binding sites; expression of the reporter genes then indicates interaction between a candidate protein and the target protein.

EXAMPLE 7: Full-length cloning of Cluster I genes

The polymerase chain reaction (PCR), which is a well known procedure for *in vitro* enzymatic amplification of a specific DNA segment, can be used for direct cloning of Protein Cluster I genes. Tissue cDNA can be amplified by PCR and cloned into an appropriate plasmid and sequenced. For reviews, see e.g. Hooft van Huijsduijnen (1998) *PCR-assisted cDNA cloning: a guided tour of the minefield.* Biotechniques 24:390-392; Lenstra (1995) *The applications of the polymerase chain reaction in the life sciences.* Cellular & Molecular Biology 41:603-614; or Rashtchian (1995) *Novel methods for cloning and engineering genes using the polymerase chain reaction.* Current Opinion in Biotechnology 6:30-36. Various methods for generating suitable ends to facilitate the direct cloning of PCR products are given e.g. in Ausubel et al. *supra* (section 15.7).

In an alternative approach to isolate a cDNA clone encoding a full length protein of Protein Cluster I, a DNA fragment corresponding to a nucleotide sequence selected from the group consisting of SEQ ID NO: 1, 3, 5 or 7, or a portion thereof, can be used as a probe for hybridization screening of a phage cDNA library. The DNA fragment is amplified by the polymerase chain reaction (PCR) method. The primers are preferably 10 to 25 nucleotides in length and are determined by procedures well known to those skilled in the art. A lambda phage library containing cDNAs cloned into lambda phage-vectors is plated on agar plates with *E. coli* host cells, and grown. Phage plaques are transferred to nylon membranes, which are hybridized with a DNA probe prepared as described above. Positive colonies are isolated from the plates. Plasmids containing cDNA are rescued from the isolated phages by standard methods. Plasmid DNA is isolated from the clones. The size of the insert is determined by digesting the plasmid with appropriate restriction enzymes. The sequence of the entire insert is determined by automated sequencing of the plasmids.

EXAMPLE 8: Recombinant expression of proteins in eukaryotic host cells

To produce proteins of Cluster I, a polypeptide-encoding nucleic acid molecule is expressed in a suitable host cell using a suitable expression vector and standard genetic engineering techniques. For example, the polypeptide-encoding sequence is subcloned into a commercial expression vector and transfected into mammalian, e.g. Chinese Hamster Ovary (CHO), cells using a standard transfection reagent. Cells stably expressing a protein are selected. Optionally, the protein may be purified from the cells using standard chromatographic techniques. To facilitate purification, antisera is raised against one or more synthetic peptide sequences that correspond to portions of the amino acid sequence, and the antisera is used to affinity purify the protein.

EXAMPLE 9: Determination of gene function

Methods are known in the art for elucidating the biological function or mode of action of individual genes. For instance, RNA interference (RNAi) offers a way of specifically and potently inactivating a cloned gene, and is proving a powerful tool for investigating gene function. For reviews, see e.g. Fire (1999) *RNA-triggered gene silencing.* Trends in Genetics 15:358-363; or Kuwabara & Coulson (2000) *RNAi–prospects for a general technique for determining gene function.* Parasitology Today 16:347-349. When double-stranded RNA (dsRNA) corresponding to a sense and antisense sequence of an endogenous mRNA is introduced into a cell, the cognate mRNA is degraded and the gene is silenced. This type of posttranscriptional gene silencing (PTGS) was first discovered in *C. elegans* (Fire et al., (1998) Nature 391:806-811). RNA interference has recently been used for targeting nearly 90% of predicted genes on *C. elegans* chromosome I (Fraser et al. (2000) Nature 408: 325-330) and 96% of predicted genes on *C. elegans* chromosome III (Gönczy et al. (2000) Nature 408:331-336).

TABLE I

Alignment of polypeptides in Protein Cluster I

```
                    *    **  ***** *****  **  ***** * **  **   * **
SEQ ID NO: 2  MESKMGELPLDINIQEPRWDQSTFLGRARHFFTVTDPRNLLLSGAQLEASRNIVQNYRAGVVTPGITEDQLWRAKY (76)
SEQ ID NO: 4  MESKMGELPLDINIQEPRWDQSTFLGRARHFFTVTDPRNLLLSGAQLEASRNIVQNY------------------- (57)
SEQ ID NO: 6  ---MEADLSG-FNIDAPRWDQRTFLGRVKHFLNITDPRTVFVSERELDWAKVMVEKSRMGVVPPGTQVEQLLYAKK (72)
SEQ ID NO: 8  ---MSGELPPNINIKEPRWDQSTFIGRANHFFTVTDPRNILLTNEQLESARKIVHDYRQGIVPPGLTENELWRAKY (73)

                                                *  * ***************
SEQ ID NO: 2  VYDSAFHPDTGEKVVLIGRMSAQVPMNMTITGCMLTFYRKTPTVVFWQWVNQSFNAIVNYSNRSGDTPITVRQLGT (152)
SEQ ID NO: 4  ------------------------------------RKTPTVVFWQWVNQSFNAIV------------------ (77)
SEQ ID NO: 6  LYDSAFHPDTGEKMNVIGRMSFQLPGGMIITGFMLQFYRTMPAVIFWQWVNQSFNALVNYTNRNAASPTSVRQMAL (148)
SEQ ID NO: 8  IYDSAFHPDTGEKMILIGRMSAQVPMNMTITGCMMTFYRTTPAVLFWQWINQSFNAVVNYTNRSGDAPLTVNELGT (149)


SEQ ID NO: 2  AYVSATTGAVATALGLKSLTKHLPPLVGRFVPFAAVAAANCINIPLMRQRELQVGIPVADEAGQRLGYSVTAAKQG (228)
SEQ ID NO: 4  ------------------------------------------------------------------------
SEQ ID NO: 6  SYFTATTTAVATAVGMNMLTKKAPPLVGRWVPFAAVAAANCVNIPMMRQRELIKGICVKDRNENEIGHSRRAAAIG (224)
SEQ ID NO: 8  AYVSATTGAVATALGLNALTKHVSPLIGRFVPFAAVAAANCINIPLMRQRELKVGIPVTDENGNRLGESANAAKQA (225)


SEQ ID NO: 2  IFQVVISRICMAIPAMAIPPLIMDTLEKKDFLK---------------------------------------------- (261)
SEQ ID NO: 4  ----------------------------------------------------------------------------------
SEQ ID NO: 6  ITQVVISRITMSAPGMILLPVIMERLEKLHFMQKVKVLHAPLQVMLSGCFLIFMVPVACGLFPQKCELPVSYLEPKLQDTIKAKYGELEPYVYFNKGL (300)
SEQ ID NO: 8  ITQVVVSRILMAAPGMAIPPFIMNTLEKKAFLKRFPWMSAPIQVGLVGFCLVFATPLCCALFPQKSSMSVTSLEAELQAKIQESHPELR-RVYFNKGL (301)
```

CLAIMS

1. An isolated nucleic acid molecule selected from:
(a) nucleic acid molecules comprising a nucleotide sequence as shown in SEQ ID NO: 1, 3, 5 or 7;
(b) nucleic acid molecules comprising a nucleotide sequence capable of hybridizing, under stringent hybridization conditions, to a nucleotide sequence complementary to the polypeptide coding region of a nucleic acid molecule as defined in (a); and
(c) nucleic acid molecules comprising a nucleic acid sequence which is degenerate as a result of the genetic code to a nucleotide sequence as defined in (a) or (b).

2. An isolated polypeptide encoded by the nucleic acid molecule according to claim 1.

3. The isolated polypeptide according to claim 2 having an amino acid sequence shown as SEQ ID NO: 2, 4, 6, or 8 in the Sequence Listing

4. A vector harboring the nucleic acid molecule according to claim 1.

5. A replicable expression vector which carries and is capable of mediating the expression of a nucleotide sequence according to claim 1.

6. A cultured host cell harboring a vector according to claim 4 or 5.

7. A process for production of a polypeptide, comprising culturing a host cell according to claim 6 under conditions whereby said polypeptide is produced, and recovering said polypeptide.

8. A method for identifying an agent capable of modulating a nucleic acid molecule according to claim 1, comprising

(i) providing a cell comprising the said nucleic acid molecule;

(ii) contacting said cell with a candidate agent; and

(iii) monitoring said cell for an effect that is not present in the absence of said candidate agent.

ABSTRACT

The present invention relates to the identification of a human gene family expressed in metabolically relevant tissues. The genes encode a group polypeptides referred to as "Protein Cluster I" which are predicted to be useful in the diagnosis of metabolic diseases, such as obesity and diabetes, as well as in the identification of agents useful in the treatment of the said diseases.

Fig. 1a

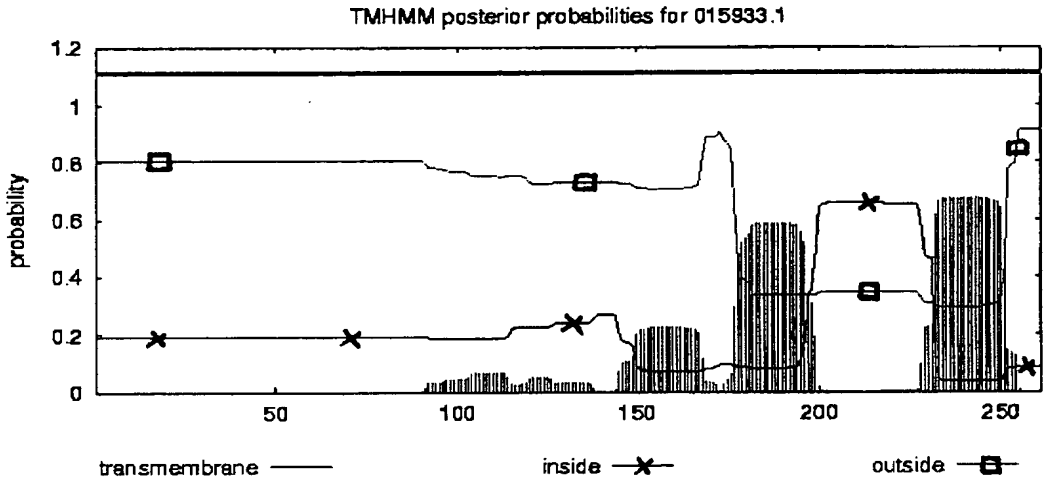

TMHMM posterior probabilities for 015933.1
probability
1.2
1
0.8
0.6
0.4
0.2
0
50
100
150
200
250
transmembrane
inside
outside

Fig. 1b

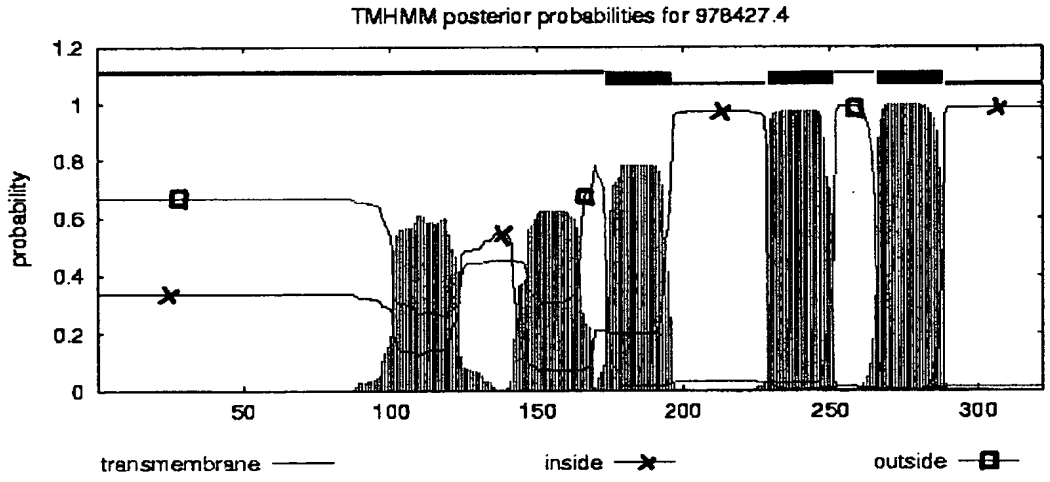
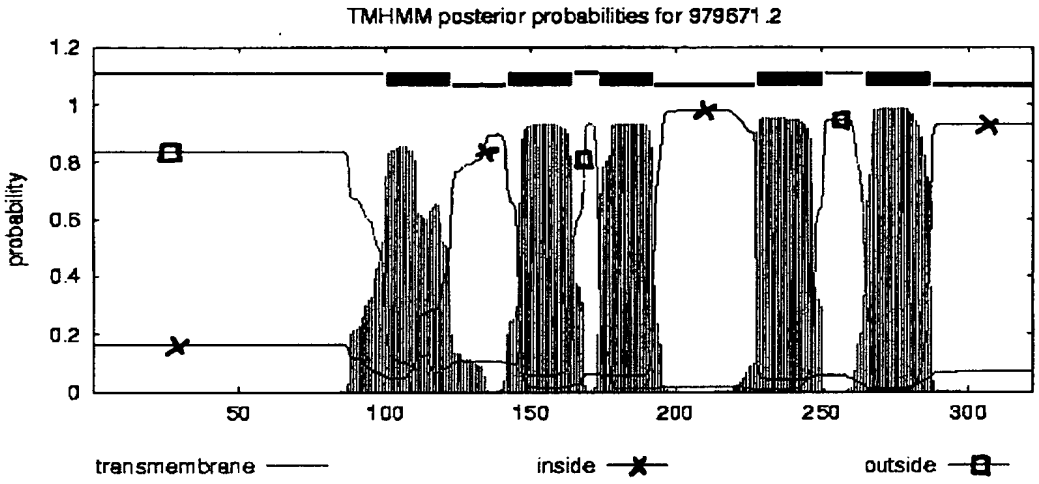

TMHMM posterior probabilities for 978427.4
probability
1.2
1
0.8
0.6
0.4
0.2
0
50
100
150
200
250
300
transmembrane
inside
outside

Fig. 1c


TMHMM posterior probabilities for 979671.2
probability
1.2
1
0.8
0.6
0.4
0.2
0
50
100
150
200
250
300
transmembrane
inside
outside

SEQUENCE LISTING

<110> Pharmacia AB

<120> Protein Cluster I

<130> 00349

<160> 8

<170> PatentIn version 3.0

<210> 1
<211> 1232
<212> DNA
<213> human

<220>
<221> CDS
<222> (450)..(1232)

<400> 1

```
cccttaggcg ccagggacag ccgagcgtta cctggtcccg ggcagcggag ttctttaccc     60

accccagttc tggttctgac gccctagctc attccgcaaa tttagggctt gggtctggct    120

tgttcccctc cggctcgaac cacctcttct ctgagccgag ccagctaccg gggctcctgg    180

aattgccacc cctccctggg caccccttgag gcctccgtgg agggacgtca cggggcagag    240

cgggacgtga gcctgagttt gctgcaggcg tgctctgtgt ggtggctggg ttctgccaat    300

ccccgtgccc accgggtggg cgcggccggg aagctcctgc ccctccctgc tggtcggcgt    360

cacgcgtgac gtcccgcgtg atggctggga gggcccggcg gcgacagcgg aggcagagag    420

gaaggcggtt ctgagagctt cagagagcg atg gaa agc aaa atg ggt gaa ttg      473
                               Met Glu Ser Lys Met Gly Glu Leu
                               1               5

cct tta gac atc aac atc cag gaa cct cgc tgg gac caa agt act ttc      521
Pro Leu Asp Ile Asn Ile Gln Glu Pro Arg Trp Asp Gln Ser Thr Phe
    10                  15                  20

ctg ggc aga gcc cgg cac ttt ttc act gtt act gat cct cga aat ctg      569
Leu Gly Arg Ala Arg His Phe Phe Thr Val Thr Asp Pro Arg Asn Leu
25                  30                  35                  40

ctg ctg tcc ggg gca cag ctg gaa gct tct cgg aac atc gtg cag aac      617
Leu Leu Ser Gly Ala Gln Leu Glu Ala Ser Arg Asn Ile Val Gln Asn
                45                  50                  55

tac agg gcc ggc gtg gtg acc cca ggg atc acc gag gac cag ctg tgg      665
Tyr Arg Ala Gly Val Val Thr Pro Gly Ile Thr Glu Asp Gln Leu Trp
            60                  65                  70
```

```
agg gcc aag tat gtg tat gac tcc gcc ttc cat ccg gac aca ggg gag      713
Arg Ala Lys Tyr Val Tyr Asp Ser Ala Phe His Pro Asp Thr Gly Glu
        75                  80                  85

aag gtg gtc ctg att ggc cgc atg tca gcc cag gtg ccc atg aac atg      761
Lys Val Val Leu Ile Gly Arg Met Ser Ala Gln Val Pro Met Asn Met
    90                  95                  100

acc atc act ggc tgc atg ctc aca ttc tac agg aag acc cca acc gtg      809
Thr Ile Thr Gly Cys Met Leu Thr Phe Tyr Arg Lys Thr Pro Thr Val
105                 110                 115                 120

gtg ttc tgg cag tgg gtg aat cag tcc ttc aat gcc att gtt aac tac      857
Val Phe Trp Gln Trp Val Asn Gln Ser Phe Asn Ala Ile Val Asn Tyr
                125                 130                 135

tcc aac cgc agt ggt gac act ccc atc act gtg agg cag ctg ggg aca      905
Ser Asn Arg Ser Gly Asp Thr Pro Ile Thr Val Arg Gln Leu Gly Thr
            140                 145                 150

gcc tat gtg agt gcc acc act gga gct gtg gcc acg gcc ctg gga ctc      953
Ala Tyr Val Ser Ala Thr Thr Gly Ala Val Ala Thr Ala Leu Gly Leu
        155                 160                 165

aaa tcc ctc acc aag cac ctg ccc ccc ttg gtc ggc aga ttt gtg ccc     1001
Lys Ser Leu Thr Lys His Leu Pro Pro Leu Val Gly Arg Phe Val Pro
    170                 175                 180

ttt gca gca gtg gca gct gcc aac tgc atc aac atc ccc ctg atg agg     1049
Phe Ala Ala Val Ala Ala Ala Asn Cys Ile Asn Ile Pro Leu Met Arg
185                 190                 195                 200

cag aga gag ctg cag gtg ggc atc ccg gtg gct gat gag gca ggt cag     1097
Gln Arg Glu Leu Gln Val Gly Ile Pro Val Ala Asp Glu Ala Gly Gln
                205                 210                 215

agg ctt ggc tac tcg gtg act gca gcc aag cag gga atc ttc cag gtg     1145
Arg Leu Gly Tyr Ser Val Thr Ala Ala Lys Gln Gly Ile Phe Gln Val
            220                 225                 230

gtg att tca aga atc tgc atg gcg att cct gcc atg gcc atc cca cca     1193
Val Ile Ser Arg Ile Cys Met Ala Ile Pro Ala Met Ala Ile Pro Pro
        235                 240                 245

ctg atc atg gac act ctg gag aag aaa gac ttc ctg aag                 1232
Leu Ile Met Asp Thr Leu Glu Lys Lys Asp Phe Leu Lys
    250                 255                 260
```

<210> 2
<211> 261
<212> PRT
<213> human

<400> 2

```
Met Glu Ser Lys Met Gly Glu Leu Pro Leu Asp Ile Asn Ile Gln Glu
1               5                   10                  15

Pro Arg Trp Asp Gln Ser Thr Phe Leu Gly Arg Ala Arg His Phe Phe
            20                  25                  30
```

```
Thr Val Thr Asp Pro Arg Asn Leu Leu Leu Ser Gly Ala Gln Leu Glu
        35                  40                  45

Ala Ser Arg Asn Ile Val Gln Asn Tyr Arg Ala Gly Val Val Thr Pro
    50                  55                  60

Gly Ile Thr Glu Asp Gln Leu Trp Arg Ala Lys Tyr Val Tyr Asp Ser
65                  70                  75                  80

Ala Phe His Pro Asp Thr Gly Glu Lys Val Val Leu Ile Gly Arg Met
                85                  90                  95

Ser Ala Gln Val Pro Met Asn Met Thr Ile Thr Gly Cys Met Leu Thr
            100                 105                 110

Phe Tyr Arg Lys Thr Pro Thr Val Val Phe Trp Gln Trp Val Asn Gln
        115                 120                 125

Ser Phe Asn Ala Ile Val Asn Tyr Ser Asn Arg Ser Gly Asp Thr Pro
    130                 135                 140

Ile Thr Val Arg Gln Leu Gly Thr Ala Tyr Val Ser Ala Thr Thr Gly
145                 150                 155                 160

Ala Val Ala Thr Ala Leu Gly Leu Lys Ser Leu Thr Lys His Leu Pro
                165                 170                 175

Pro Leu Val Gly Arg Phe Val Pro Phe Ala Ala Val Ala Ala Ala Asn
            180                 185                 190

Cys Ile Asn Ile Pro Leu Met Arg Gln Arg Glu Leu Gln Val Gly Ile
        195                 200                 205

Pro Val Ala Asp Glu Ala Gly Gln Arg Leu Gly Tyr Ser Val Thr Ala
    210                 215                 220

Ala Lys Gln Gly Ile Phe Gln Val Val Ile Ser Arg Ile Cys Met Ala
225                 230                 235                 240

Ile Pro Ala Met Ala Ile Pro Pro Leu Ile Met Asp Thr Leu Glu Lys
                245                 250                 255

Lys Asp Phe Leu Lys
            260
```

<210> 3
<211> 1061
<212> DNA
<213> human

<220>
<221> CDS
<222> (450)..(680)

<400> 3
ccctttaggcg ccagggacag ccgagcgtta cctggtcccg ggcagcggag ttctttaccc 60

accccagttc tggttctgac gccctagctc attccgcaaa tttagggctt gggtctggct 120

```
tgttcccctc cggctcgaac cacctcttct ctgagccgag ccagctaccg gggctcctgg    180

aattgccacc cctccctggg caccettgag gcctccgtgg agggacgtca cggggcagag    240

cgggacgtga gcctgagttt gctgcaggcg tgctctgtgt ggtggctggg ttctgccaat    300

ccccgtgccc accgggtggg cgcggccggg aagctcctgc ccctccctgc tggtcggcgt    360

cacgcgtgac gtcccgcgtg atggctggga gggcccggcg gcgacagcgg aggcagagag    420

gaaggcggtt ctgagagctt cagagagcg atg gaa agc aaa atg ggt gaa ttg      473
                               Met Glu Ser Lys Met Gly Glu Leu
                               1               5

cct tta gac atc aac atc cag gaa cct cgc tgg gac caa agt act ttc      521
Pro Leu Asp Ile Asn Ile Gln Glu Pro Arg Trp Asp Gln Ser Thr Phe
    10                  15                  20

ctg ggc aga gcc cgg cac ttt ttc act gtt act gat cct cga aat ctg      569
Leu Gly Arg Ala Arg His Phe Phe Thr Val Thr Asp Pro Arg Asn Leu
25                  30                  35                  40

ctg ctg tcc ggg gca cag ctg gaa gct tct cgg aac atc gtg cag aac      617
Leu Leu Ser Gly Ala Gln Leu Glu Ala Ser Arg Asn Ile Val Gln Asn
                45                  50                  55

tac agg aag acc cca acc gtg gtg ttc tgg cag tgg gtg aat cag tcc      665
Tyr Arg Lys Thr Pro Thr Val Val Phe Trp Gln Trp Val Asn Gln Ser
            60                  65                  70

ttc aat gcc att gtt aactactcca accgcagtgg tgacactccc atcactgtga      720
Phe Asn Ala Ile Val
        75

ggcagctggg gacagcctat gtgagtgcca ccactggagc tgtggccacg gccctgggac    780

tcaaatccct caccaagcac ctgccccccc tggtcggcag atttgtgccc tttgcagcag    840

tggcagctgc caactgcatc aacatccccc tgatgaggca gagagagctg caggtgggca    900

tcccggtggc tgatgaggca ggtcagaggc ttggctactc ggtgactgca gccaagcagg    960

gaatcttcca ggtggtgatt tcaagaatct gcatggcgat tcctgccatg gccatcccac   1020

cactgatcat ggacactctg gagaagaaag acttcctgaa g                       1061


<210>  4
<211>  77
<212>  PRT
<213>  human

<400>  4
Met Glu Ser Lys Met Gly Glu Leu Pro Leu Asp Ile Asn Ile Gln Glu
1               5                   10                  15

Pro Arg Trp Asp Gln Ser Thr Phe Leu Gly Arg Ala Arg His Phe Phe
            20                  25                  30
```

```
Thr Val Thr Asp Pro Arg Asn Leu Leu Leu Ser Gly Ala Gln Leu Glu
        35                  40                  45

Ala Ser Arg Asn Ile Val Gln Asn Tyr Arg Lys Thr Pro Thr Val Val
    50                  55                  60

Phe Trp Gln Trp Val Asn Gln Ser Phe Asn Ala Ile Val
65                  70                  75


<210>  5
<211>  1567
<212>  DNA
<213>  human

<220>
<221>  CDS
<222>  (47)..(1015)

<400>  5
gggcatttgt cccgggacca ggtccacagt tttatgtgtg agcaag atg gag gct        55
                                                  Met Glu Ala
                                                  1

gac ctg tct ggc ttt aac atc gat gcc ccc cgt tgg gac cag cgc acc       103
Asp Leu Ser Gly Phe Asn Ile Asp Ala Pro Arg Trp Asp Gln Arg Thr
    5                   10                  15

ttc ctg ggg aga gtg aag cac ttc cta aac atc acg gac ccc cgc act       151
Phe Leu Gly Arg Val Lys His Phe Leu Asn Ile Thr Asp Pro Arg Thr
20                  25                  30                  35

gtc ttt gta tct gag cgg gag ctg gac tgg gcc aag gtg atg gtg gag       199
Val Phe Val Ser Glu Arg Glu Leu Asp Trp Ala Lys Val Met Val Glu
                40                  45                  50

aag agc agg atg ggg gtt gtg ccc cca ggc acc caa gtg gag cag ctg       247
Lys Ser Arg Met Gly Val Val Pro Pro Gly Thr Gln Val Glu Gln Leu
            55                  60                  65

ctg tat gcc aag aag ctg tat gac tcg gcc ttc cac ccc gac act ggg       295
Leu Tyr Ala Lys Lys Leu Tyr Asp Ser Ala Phe His Pro Asp Thr Gly
        70                  75                  80

gag aag atg aat gtc atc ggg cgc atg tct ttc cag ctt cct ggc ggc       343
Glu Lys Met Asn Val Ile Gly Arg Met Ser Phe Gln Leu Pro Gly Gly
    85                  90                  95

atg atc atc acg ggc ttc atg ctc cag ttc tac agg acg atg ccg gcg       391
Met Ile Ile Thr Gly Phe Met Leu Gln Phe Tyr Arg Thr Met Pro Ala
100                 105                 110                 115

gtg atc ttc tgg cag tgg gtg aac cag tcc ttc aat gcc tta gtc aac       439
Val Ile Phe Trp Gln Trp Val Asn Gln Ser Phe Asn Ala Leu Val Asn
                120                 125                 130

tac acc aac agg aat gcg gct tcc ccc aca tca gtc agg cag atg gcc       487
Tyr Thr Asn Arg Asn Ala Ala Ser Pro Thr Ser Val Arg Gln Met Ala
            135                 140                 145
```

```
ctt tcc tac ttc aca gcc aca acc act gct gtg gcc acg gct gtg ggc      535
Leu Ser Tyr Phe Thr Ala Thr Thr Thr Ala Val Ala Thr Ala Val Gly
        150                 155                 160

atg aac atg ttg aca aag aaa gcg ccg ccc ttg gtg ggc cgc tgg gtg      583
Met Asn Met Leu Thr Lys Lys Ala Pro Pro Leu Val Gly Arg Trp Val
    165                 170                 175

ccc ttt gcc gct gtg gct gcg gct aac tgt gtc aat atc ccc atg atg      631
Pro Phe Ala Ala Val Ala Ala Ala Asn Cys Val Asn Ile Pro Met Met
180                 185                 190                 195

cga cag agg gag ctc ata aag gga atc tgc gtg aag gac agg aat gaa      679
Arg Gln Arg Glu Leu Ile Lys Gly Ile Cys Val Lys Asp Arg Asn Glu
                200                 205                 210

aat gag att ggt cat tcc cgg aga gct gcg gcc ata ggc atc acc caa      727
Asn Glu Ile Gly His Ser Arg Arg Ala Ala Ala Ile Gly Ile Thr Gln
            215                 220                 225

gta gtt att tct cgg atc acc atg tca gct cct ggg atg atc ttg ctg      775
Val Val Ile Ser Arg Ile Thr Met Ser Ala Pro Gly Met Ile Leu Leu
        230                 235                 240

cca gtc atc atg gaa agg ctt gag aaa ttg cac ttc atg cag aaa gtc      823
Pro Val Ile Met Glu Arg Leu Glu Lys Leu His Phe Met Gln Lys Val
    245                 250                 255

aag gtc ctg cac gcc cca ttg cag gtc atg ctg agc ggg tgc ttc ctc      871
Lys Val Leu His Ala Pro Leu Gln Val Met Leu Ser Gly Cys Phe Leu
260                 265                 270                 275

atc ttc atg gtg cca gtg gcg tgt ggg ctt ttc cca cag aaa tgt gaa      919
Ile Phe Met Val Pro Val Ala Cys Gly Leu Phe Pro Gln Lys Cys Glu
                280                 285                 290

ttg cca gtt tcc tat ctg gaa ccg aag ctc caa gac act atc aag gcc      967
Leu Pro Val Ser Tyr Leu Glu Pro Lys Leu Gln Asp Thr Ile Lys Ala
            295                 300                 305

aag tat gga gaa ctt gag cct tat gtc tac ttc aat aag ggt ctc taa     1015
Lys Tyr Gly Glu Leu Glu Pro Tyr Val Tyr Phe Asn Lys Gly Leu
        310                 315                 320

atgccccact tcagcaagga ccagtctatt cccatattca ccagctcctc cttagctacg   1075

tgcacacttg tgtcctcctt ccccttttgcc aacaaggcct gaaggccagg gtagattggg   1135

gggtgggaca atgaatgcct catacttaca ccctggtact ggttgattgg acctcagggg   1195

aaaaaagtga aaaagggtag caaaggccaa tgtcttctag ctgcttcctc aacccctgtc   1255

ccctgagaga ccagaagctg aggccctctc agggaggaga catccaagca aatcatttgg   1315

aaaagttagg aaacctttag gattctggtt ccagccaggg ttgaggaaaa gaccttggat   1375

caaaaggaag cttctatacc tctttcttct tcgcttcctc ctctcccaag caatggaaac   1435

ttttacccat gtaattctag ctgaactcag gaaaaagaag ggggaaagga ctctgtcccc   1495
```

ttggggctca tcacccttcc acatcctcct cctcgtagcc ccctggtcag gcagcttctt 1555

tttttttttt tc 1567

<210> 6
<211> 322
<212> PRT
<213> human

<400> 6

Met Glu Ala Asp Leu Ser Gly Phe Asn Ile Asp Ala Pro Arg Trp Asp
1 5 10 15

Gln Arg Thr Phe Leu Gly Arg Val Lys His Phe Leu Asn Ile Thr Asp
20 25 30

Pro Arg Thr Val Phe Val Ser Glu Arg Glu Leu Asp Trp Ala Lys Val
35 40 45

Met Val Glu Lys Ser Arg Met Gly Val Val Pro Pro Gly Thr Gln Val
50 55 60

Glu Gln Leu Leu Tyr Ala Lys Lys Leu Tyr Asp Ser Ala Phe His Pro
65 70 75 80

Asp Thr Gly Glu Lys Met Asn Val Ile Gly Arg Met Ser Phe Gln Leu
85 90 95

Pro Gly Gly Met Ile Ile Thr Gly Phe Met Leu Gln Phe Tyr Arg Thr
100 105 110

Met Pro Ala Val Ile Phe Trp Gln Trp Val Asn Gln Ser Phe Asn Ala
115 120 125

Leu Val Asn Tyr Thr Asn Arg Asn Ala Ala Ser Pro Thr Ser Val Arg
130 135 140

Gln Met Ala Leu Ser Tyr Phe Thr Ala Thr Thr Thr Ala Val Ala Thr
145 150 155 160

Ala Val Gly Met Asn Met Leu Thr Lys Lys Ala Pro Pro Leu Val Gly
165 170 175

Arg Trp Val Pro Phe Ala Ala Val Ala Ala Ala Asn Cys Val Asn Ile
180 185 190

Pro Met Met Arg Gln Arg Glu Leu Ile Lys Gly Ile Cys Val Lys Asp
195 200 205

Arg Asn Glu Asn Glu Ile Gly His Ser Arg Arg Ala Ala Ala Ile Gly
210 215 220

Ile Thr Gln Val Val Ile Ser Arg Ile Thr Met Ser Ala Pro Gly Met
225 230 235 240

Ile Leu Leu Pro Val Ile Met Glu Arg Leu Glu Lys Leu His Phe Met
245 250 255

```
Gln Lys Val Lys Val Leu His Ala Pro Leu Gln Val Met Leu Ser Gly
            260                 265                 270

Cys Phe Leu Ile Phe Met Val Pro Val Ala Cys Gly Leu Phe Pro Gln
        275                 280                 285

Lys Cys Glu Leu Pro Val Ser Tyr Leu Glu Pro Lys Leu Gln Asp Thr
    290                 295                 300

Ile Lys Ala Lys Tyr Gly Glu Leu Glu Pro Tyr Val Tyr Phe Asn Lys
305                 310                 315                 320

Gly Leu
```

<210> 7
<211> 2269
<212> DNA
<213> human

<220>
<221> CDS
<222> (125)..(1093)

<220>
<221> misc_feature
<222> (25)..(25)
<223> n=A,T,G or C

<400> 7

```
gacgcgctcc ggggacgcgc gaggncgccg tggcgggaga agcgtttccg gtggcggcgg     60

aggctgcact gagcgggacc tggcgagcag cgcgggcggc agcccggggg aagcgtccgg    120

gacc atg tct gga gaa cta cca cca aac att aac atc aag gaa cct cga      169
     Met Ser Gly Glu Leu Pro Pro Asn Ile Asn Ile Lys Glu Pro Arg
     1               5                   10                  15

tgg gat caa agc act ttc att gga cga gcc aat cat ttc ttc act gta      217
Trp Asp Gln Ser Thr Phe Ile Gly Arg Ala Asn His Phe Phe Thr Val
                20                  25                  30

act gac ccc agg aac att ctg tta acc aac gaa caa ctc gag agt gcg      265
Thr Asp Pro Arg Asn Ile Leu Leu Thr Asn Glu Gln Leu Glu Ser Ala
            35                  40                  45

aga aaa ata gta cat gat tac agg cag gga att gtt cct cct ggt ctt      313
Arg Lys Ile Val His Asp Tyr Arg Gln Gly Ile Val Pro Pro Gly Leu
        50                  55                  60

aca gaa aat gaa ttg tgg aga gca aag tac atc tat gat tca gct ttt      361
Thr Glu Asn Glu Leu Trp Arg Ala Lys Tyr Ile Tyr Asp Ser Ala Phe
    65                  70                  75

cat cct gac act ggt gag aag atg att ttg ata gga aga atg tca gcc      409
His Pro Asp Thr Gly Glu Lys Met Ile Leu Ile Gly Arg Met Ser Ala
80                  85                  90                  95
```

```
cag gtt ccc atg aac atg acc atc aca ggt tgt atg atg acg ttt tac      457
Gln Val Pro Met Asn Met Thr Ile Thr Gly Cys Met Met Thr Phe Tyr
                100                 105                 110

agg act acg ccg gct gtg ctg ttc tgg cag tgg att aac cag tcc ttc      505
Arg Thr Thr Pro Ala Val Leu Phe Trp Gln Trp Ile Asn Gln Ser Phe
            115                 120                 125

aat gcc gtc gtc aat tac acc aac aga agt gga gac gca ccc ctc act      553
Asn Ala Val Val Asn Tyr Thr Asn Arg Ser Gly Asp Ala Pro Leu Thr
        130                 135                 140

gtc aat gag ttg gga aca gct tac gtt tct gca aca act ggt gcc gta      601
Val Asn Glu Leu Gly Thr Ala Tyr Val Ser Ala Thr Thr Gly Ala Val
    145                 150                 155

gca aca gct cta gga ctc aat gca ttg acc aag cat gtc tca cca ctg      649
Ala Thr Ala Leu Gly Leu Asn Ala Leu Thr Lys His Val Ser Pro Leu
160                 165                 170                 175

ata gga cgt ttt gtt ccc ttt gct gcc gta gct gct gct aat tgc att      697
Ile Gly Arg Phe Val Pro Phe Ala Ala Val Ala Ala Ala Asn Cys Ile
                180                 185                 190

aat att cca tta atg agg caa agg gaa ctc aaa gtt ggc att ccc gtc      745
Asn Ile Pro Leu Met Arg Gln Arg Glu Leu Lys Val Gly Ile Pro Val
            195                 200                 205

acg gat gag aat ggg aac cgc ttg ggg gag tcg gcg aac gct gcg aaa      793
Thr Asp Glu Asn Gly Asn Arg Leu Gly Glu Ser Ala Asn Ala Ala Lys
        210                 215                 220

caa gcc atc acg caa gtt gtc gtg tcc agg att ctc atg gca gcc cct      841
Gln Ala Ile Thr Gln Val Val Val Ser Arg Ile Leu Met Ala Ala Pro
    225                 230                 235

ggc atg gcc atc cct cca ttc att atg aac act ttg gaa aag aaa gcc      889
Gly Met Ala Ile Pro Pro Phe Ile Met Asn Thr Leu Glu Lys Lys Ala
240                 245                 250                 255

ttt ttg aag agg ttc cca tgg atg agt gca ccc att caa gtt ggg tta      937
Phe Leu Lys Arg Phe Pro Trp Met Ser Ala Pro Ile Gln Val Gly Leu
                260                 265                 270

gtt ggc ttc tgt ttg gtg ttt gct aca ccc ctg tgt tgt gcc ctg ttt      985
Val Gly Phe Cys Leu Val Phe Ala Thr Pro Leu Cys Cys Ala Leu Phe
            275                 280                 285

cct cag aaa agt tcc atg tct gtg aca agc ttg gag gcc gag ttg caa     1033
Pro Gln Lys Ser Ser Met Ser Val Thr Ser Leu Glu Ala Glu Leu Gln
        290                 295                 300

gct aag atc caa gag agc cat cct gaa ttg cga cgc gtg tac ttc aat     1081
Ala Lys Ile Gln Glu Ser His Pro Glu Leu Arg Arg Val Tyr Phe Asn
    305                 310                 315

aag gga ttg taa agcagagagg aaacctctgc agctcattct gccactgcaa         1133
Lys Gly Leu
320
```

agctggtgta gccatgctgg tgagaaaaat cctgttcaac ctgggttctc ccagttacgg 1193

aaacctttta aagatccaca ttagcctttt agaataaagc tgctacttta acagagcacc 1253

tggcgtgggc caagtgcctg atactccctt acactgaatc atgttatgat ttatagaaat 1313

acctttcctg tagcttttat agtcattgtt tttcaaagac gatataccag ccctcaccca 1373

ggttttaaaa aagcactggt aggcatagaa taggtgctca gtatatggtc agtaaatgtt 1433

ctattgatta tcaatcagtg aaaaaagaaa tctgtttaaa atactgaatt ttcatctcac 1493

tcccattgca aatcaaggag atctcagcag tgaactggga aaatacaaaa gctctgggct 1553

aatctataaa aacttacctg aaatattaag ggcagtttgc ttctagtttg gggattgcgc 1613

tagcccaatg aaggtgatga agcttttgga tttggagggt aaaagctcct tcacacccct 1673

tccaaaagtc agtcacagac cactgcaaca tgccttccct gctggatcat tatatacatt 1733

cagattgtga gtggattgcc ttggttgact tttaatttat tgtttttgt tcttataaag 1793

atgataatct taccttgcag ttattgactt tatattcaat tatttacatc aaataatgaa 1853

ataactgaaa tgtacaaatg tcaaattttg gaagtatatt caataccaat gctgtatgag 1913

tgggctgaat ccagttcatt gttttttttt tggtaagaag tgagactaca gttccagcta 1973

cctacatgtc ttttcttgtc atccttatag atctctttgg ctttcagaaa gatacagtga 2033

taatgtgtgt atgaatcagt cacaatgaat tttacttgaa tattgtatgt tgcattccac 2093

ttcatttgaa aataatgaaa ccatgtacca ctgtttacat catctgtagt gatttcatag 2153

ataatatatt taatatgaca gattatgttt caactctgta gatgtttaac gtcatagaca 2213

gtcggccctc tgtatccgtg agctctatat ctgtgaattc aaccaagttt ggatgg 2269

<210> 8
<211> 322
<212> PRT
<213> human

<220>
<221> misc_feature
<222> (25)..(25)
<223> n=A,T,G or C

<400> 8

Met Ser Gly Glu Leu Pro Pro Asn Ile Asn Ile Lys Glu Pro Arg Trp
1 5 10 15

Asp Gln Ser Thr Phe Ile Gly Arg Ala Asn His Phe Phe Thr Val Thr
20 25 30

Asp Pro Arg Asn Ile Leu Leu Thr Asn Glu Gln Leu Glu Ser Ala Arg
35 40 45

Lys Ile Val His Asp Tyr Arg Gln Gly Ile Val Pro Pro Gly Leu Thr
50 55 60

Glu Asn Glu Leu Trp Arg Ala Lys Tyr Ile Tyr Asp Ser Ala Phe His
65 70 75 80

Pro Asp Thr Gly Glu Lys Met Ile Leu Ile Gly Arg Met Ser Ala Gln
85 90 95

Val Pro Met Asn Met Thr Ile Thr Gly Cys Met Met Thr Phe Tyr Arg
100 105 110

Thr Thr Pro Ala Val Leu Phe Trp Gln Trp Ile Asn Gln Ser Phe Asn
115 120 125

Ala Val Val Asn Tyr Thr Asn Arg Ser Gly Asp Ala Pro Leu Thr Val
130 135 140

Asn Glu Leu Gly Thr Ala Tyr Val Ser Ala Thr Thr Gly Ala Val Ala
145 150 155 160

Thr Ala Leu Gly Leu Asn Ala Leu Thr Lys His Val Ser Pro Leu Ile
165 170 175

Gly Arg Phe Val Pro Phe Ala Ala Val Ala Ala Ala Asn Cys Ile Asn
180 185 190

Ile Pro Leu Met Arg Gln Arg Glu Leu Lys Val Gly Ile Pro Val Thr
195 200 205

Asp Glu Asn Gly Asn Arg Leu Gly Glu Ser Ala Asn Ala Ala Lys Gln
210 215 220

Ala Ile Thr Gln Val Val Val Ser Arg Ile Leu Met Ala Ala Pro Gly
225 230 235 240

Met Ala Ile Pro Pro Phe Ile Met Asn Thr Leu Glu Lys Lys Ala Phe
245 250 255

Leu Lys Arg Phe Pro Trp Met Ser Ala Pro Ile Gln Val Gly Leu Val
260 265 270

Gly Phe Cys Leu Val Phe Ala Thr Pro Leu Cys Cys Ala Leu Phe Pro
275 280 285

Gln Lys Ser Ser Met Ser Val Thr Ser Leu Glu Ala Glu Leu Gln Ala
290 295 300

Lys Ile Gln Glu Ser His Pro Glu Leu Arg Arg Val Tyr Phe Asn Lys
305 310 315 320

Gly Leu